From: PETERSON Jenn L
To: Eric Blischke/R10/USEPA/US@EPA
Cc: Burt Shephard/R10/USEPA/US@EPA; Robert W. Gensemer
Subject: RE: Bioassay Interpretation
Date: 06/12/2008 12:38 PM

I realize this is a difficult topic to reach agreement on, and I am not trying to make this harder :-).

I am glad we seem to agree that the interpretation of the empirical results would be using the 10, 20 and 30 thresholds. So, my next question is what is the primary purpose of the model? Right now it was my assumption that it was for the risk assessment - to predict those same thresholds where we have chemistry but not toxicity data. Now, it is true that we can predict these thresholds with a varying degree of confidence, and that is where reliability comes in. How you calculate uncertainty in predictions has a great deal to do with your objectives. I outlined what I though our objectives should be for the risk assessment my e-mail while back on reliability. However, I thought the next iteration of the risk assessment would lay out model predictions for each endpoint and the associated reliability. We could then build from than to reach concurrence on how / if these model predictions could be used for SQG / PRG development. When we reach this point, it may be clear that the model does not predict some endpoints with enough confidence (e.g. maybe Hy growth). However, I have not seen this evaluation for the FPM. The LRM does seem to be working with this endpoint. If we feel it is important we can also bring in new data to improve the FPM model as Jay did with the LRM. We also haven't reach agreement on how to incorporate PAHs into the model. We did not reach agreement on these things after the round 2 report.

We may find we can predict some thresholds better than others. For example, the higher thresholds (larger magnitude responses) by nature will be easier to predict with the models. However, the potential inability to predict (if this is the case) should not be criteria for determining if an endpoint or threshold is relevant. It just means that we can get all the way there using a model and need to rely on empirical observations.

Also, if I read this correctly, you aren't just proposing to change the thresholds in some cases, but are also proposing to combine them. The end result using the RSET approach (e.g. one hit/ two hit) would be that the response of both organisms are not accounted for equally - esp. where they may differ. Or, the response of one organism would have to be at a larger magnitude to qualify as a hit. I thought we agreed that this wasn't appropriate for the risk assessment.

I think we will have to develop our own decision criteria for benthic risk. It may or may not be similar to RSET's approach, and hopefully it will involve more lines of evidence. However, I think we need the time, and presentation of data and risk assessment to make an informed decision. How does limiting the model application limit us in the future in developing SQGs or a benthic decision framework?

I still think we should leave thresholds the same for both the model and empirical data. To address their concerns to would add a higher bin to the Hy growth, and evaluate with the next report.

-Jennifer

-----Original Message-----
From: Blischke.Eric@epamail.epa.gov
[mailto:Blischke.Eric@epamail.epa.gov]
Sent: Thursday, June 12, 2008 9:30 AM
To: PETERSON Jenn L
Cc: Shephard.Burt@epamail.epa.gov; Robert W. Gensemer
Subject: RE: Bioassay Interpretation

I am just trying to chart a course here. My understanding was that we agreed to not use the RSET approach for the empirical evaluation. For the empirical data, we are leaving the endpoint thresholds at the 10%, 20% and 30% thresholds as well as flagging anything statisically signifcantly different than control as a hit. The reason for adjusting the modeling threshold is to (hopefully) be able to tease out better chemicals that likely posing risk and being consistent with future RSET evaluation criteria. I agree that this is quite a change from what we had been talking about but I believe that this is a reasonable way to go and takes maximum advantage of the the empirical data. Recall that that the hyalella growth endpoint correlates with no chemical. What is the harm in changing the threshold?

Eric

```
             "PETERSON Jenn
             L"
             <PETERSON.Jenn@d                                        To
             eq.state.or.us>          Eric Blischke/R10/USEPA/US@EPA,
                                      Burt Shephard/R10/USEPA/US@EPA,
             06/12/2008 08:45          "Robert W. Gensemer"
             AM                       <rgensemer@parametrix.com>
                                                                     cc
```

Subject
RE: Bioassay Interpretation

I am confused - I thought we had already agreed not to use RSET criteria? What is met by "using RSET criteria"? Using the framework would result in many changes to the problem formulation in addition to just changing the hit/no-hit thresholds and adding pooled decision criteria. By using the one-hit / two hit rule we are losing information. These species are exposed differently (one as infauna and sediment ingestion and the other as an epibenthic organism). I thought we had agreed that looking at both separately are important lines of evidence in evaluating risk to the benthic community. By pooling, we lose a lot of information, and we may also not gain any reliability. Combining the two species responses may further confound the model - we found that was the case when examining Washington's criteria.

I feel strongly that the chironomus growth endpoint thresholds should be left as they were. This is a 10-day test (short term) and presenting the 10% level is important. There is no justification for removing the lower threshold, 10% difference is a part of Washington's criteria and RSET's.

IF we think there is a problem with model predictions with one of the endpoints, like the HY 28 day growth, then that one alone should be evaluated further. However, I think there should be evidence that this is the case. I have not been convinced been convinced this is the case for HY growth, as I have not seen the FPM model for this endpoint. Our FPM could also benefit from RSET's new model, which brings in other regional data to improve predictibility.

Model predictions should be geared toward what we think is acceptable difference - or at least present the range (e.g. minor, moderate, severe). I am not sure why we would lose this information in the models when we think it is appropriate to evaluate empirically. We are trying to predict stations where we only have chemistry and not bioassay data - there are a lot of these stations. Why wouldn't we want this same information available?

Why is one of the stated objectives to predict a "higher threshold"? This approach minimizes false positives but drastically increases false negatives. When chemistry is the only line of evidence, this increases the likelihood that we would call a station a "no hit" when in reality it actually was. Certainly we need to balance these two, mostly for SQG / PRG development, but if we make requirement too high to call a station a hit, we will not be investigating the subsequent "no hits" further. The "investigation" should be done in the context of multiple lines of evidence in the risk assessment. I also thought the models were developed at this point to help inform the risk assessment, not just PRGs. I think decision criteria applied to the empirical or benthic models for the purposes of setting SQGs / PRGs is a separate topic. I think our goal should be to predict a relevant range of effects on the benthic community - from minor to severe. That has been the goal of all work on the project so far.

If some compromise must be made then I would advocate just changing the HY-growth bins (which I thought was the only issue remaining). By making changes to the other thresholds as well as instituting the RSET interpretive / decision making rules we are losing too much. Save the rest of the discussion for PRG development.

-Jennifer

-----Original Message-----
From: Blischke.Eric@epamail.epa.gov [
mailto:Blischke.Eric@epamail.epa.gov]
Sent: Wednesday, June 11, 2008 5:33 PM
To: PETERSON Jenn L; Shephard.Burt@epamail.epa.gov;
rgensemer@parametrix.com
Cc: Humphrey.Chip@epamail.epa.gov
Subject: Bioassay Interpretation

I just got off the phone with John Toll. This is what we tentatively agreed to:

For the empirical data:

Anything statistically different than control will be flagged as a hit. Hits will be color coded as to magnitude of the hit (10%, 20%, 30%). This will be done for each of the four endpoints (CH10 survival, CH10 growth, HY28 survival, and HY28 growth)

For the predictive models:

We will apply the RSET approach to the model development. The following criteria will be used: CH10 and HY28 survival: SL1 = 10%, SL2 = 20%; CH10 growth: SL1=20%, SL2 = 30%; HY28 growth: SL1 = 25%, SL2 = 40%. As with RSET, if any one of the four SL2 values are exceeded or any 2 of the SL1 values are exceeded, the data will be presented as a hit. Empirical data will be presented against these criteria as well.

The rationale for this approach is:

1) The empirical data collected at the PH site is of high quality and is our strongest line of evidence. This approach ensures that we will get the most out of the empirical data set.

2) It is difficult to tease out hits with the predictive models. This approach provides a higher threshold for documenting a hit in the predictive models and is consistent with the RSET approach.

John is running this by the LWG team. I am running this by you three. Please let me know your view on this approach at your earliest convenience.

Thanks, Eric